Towards Achieving Optimal Performance using Stacked Generalization Algorithm: A Case Study of Clinical Diagnosis of Malaria Fever

Abiodun Oguntimilehin[1], Olusola Adetunmbi[2], and Innocent Osho[3]
[1]Department of Computer Science, Afe Babalola University, Ado-Ekiti
[2]Department of Computer Science, Federal University of Technology, Akure
[3]Department of Animal Production and Health, Federal University of Technology, Akure

Abstract: *The birth of data mining has been a blessing to all fields of endeavours and there are numerous data mining algorithms available today. One of the major problems of mining data is the selection of the appropriate algorithm or model for a job at hand; this has led to different comparison experiments by researchers. Stacked Generalization is one of the methods of combining multiple models to give a better accuracy. The method has been investigated to be effective by many researchers over the years. This study investigates how optimal performance could be achieved using Stacked Generalization algorithm. Six different data mining algorithms (PART, REP Tree, J48, Random Tree, RIDOR and JRIP) arranged in two different orders were used as base learners to two different Meta Learners (Random Forest and NNGE) independently and the results obtained were compared in terms of classification accuracy. The study shows that the order of arrangement of the base learners and the choice of Meta Learner could affect the accuracy of the Stacked Generalization method; NNGE outperforms Random Forest as a Meta-Learner and its performance is independent of the order of arrangement of the base learners as against Random Forest. Malaria fever datasets collected from reputable hospitals in Ado-Ekiti, Ekiti State, Nigeria were purposefully used for this study because malaria is one of the major diseases killing almost a million people yearly in the tropical region of Africa, so a more accurate malaria fever diagnosis model is as well proposed as a result of this study.*

Keywords: *Data mining, ensemble learning, stacked generalization, malaria, diagnosis.*

Received October 13, 2016; accepted October 11, 2017

1. Introduction

Data mining came into existence in the 1990s and since then has gotten a popular weight in business, industry and press communities. Data mining is today synonymous with knowledge discovery and the processes of mining data are in correlated refined phases. Predictive and descriptive are the two major divisions of data mining. Predictive data mining targets prediction of a particular decision attribute or class based on the received training while descriptive creates groups from the entire dataset subject to peculiar similarities. In order words, members of each group possess similar characteristics that distinguish them from the other groups. There are numerous data mining algorithms and data mining tools available today and they have been proved to be very useful in decision making. Classification, regression, time series analysis and prediction are examples of predictive data mining while association rules, summarization, clustering and sequential pattern analysis are examples of descriptive data mining [9].

As there are numerous data mining algorithms and data mining tools today, selection of appropriate model or algorithm suitable for their needs by the end users has been a great problem. At times, end users end up selecting an algorithm not suitable for the problem. This necessitates the use of ensemble learning methods in data mining [22]. Ensemble learning is a data mining approach in which multiple learners called base learners or level learners combined their outputs to give a single better predictive model. There are two major tasks involved in the construction of an ensemble classifier. The first task is to generate two or more level learners and after which the second task of how to combine the predictive outputs of these level learners to give rise to a single stronger classifier arises [20]. Stacked Generalization is one of the methods of combining the level learners. In Stacked Generalization or Stacking, a data mining algorithm is used to combine the outputs of the base learners. Here outputs of the other data mining algorithms (base learners) are used as a training set to another data mining algorithm [1, 19, 20]. The data mining algorithm used to combine and learn the behaviours of the base learners is called a Meta-learner [20]. Many fields of endeavours today are being revolutionized by computer technology. Human expertise is now being captured and made available at reduced cost [13]. Computer technology has been successfully applied in medical field over the years to carry out diagnosis and treatment in the form of medical decision support

systems and this practice is fast increasing daily in different areas of medical problems [8]. A correct medical diagnosis will surely ensure correct treatment of the diagnosed disease or illness [23].

Malaria is a disease caused by Eukaryotic protist of the genus plasmodium and is being spread by malaria parasite carrying mosquitoes. Malaria is prevalent in Tropical region of Africa and there is a report of 500 million cases yearly [10]. This work investigates how optimality can be achieved while using Stacked Generalization and as well proposes a better method of diagnosing malaria fever.

1.1. Organization of the Paper

The body of this paper is divided into five sections. It starts with introduction as section 1; follows by section 2 which focuses literature review. The research method is discussed in section 3; following this is section 4 which presented the experimental setup and results while conclusion forms the section 5.

2. Literature Review

David et al did a comparative evaluation of Combiner and Stacked Generalization in [5]. Dichotomiser3 (ID3), Classification and Regression Tree (CART) and Baye's theorem (BAYES) classifiers were used as base classifiers. Two molecular biology sequence analysis datasets obtained from UCI machine learning database were used to train the models. Different combinations of the tree base classifiers and one meta-learner were experimented on the two different datasets. It was observed from the results that both combiner and stacked generalization performed better than the base learners. The accuracies of the two methods are almost the same but combiner is considered better in terms of accuracy and cost.

Ting and Witten [17] investigated the use of Stacked Generalization in combining models gotten from different subsets of a training set using a single learning algorithm and later different learning algorithms. Bootstrap samples called 'bagging' and disjoint samples called 'dagging' were used to train the base learners and Stacked Generalization was used to combine the models generated by the base learners. Two new methods called bag-stacking and dag-stacking evolved from this experiment. The work showed that dag-stacking and bag-stacking can also work very well for a classification problem where only small sample of the whole dataset is available or used. The work further demonstrated that bagging and bag-stacking as well as dagging and dag-stacking perform well on both stable and unstable learning algorithms. Bag-stacking as well as dag-stacking almost performs better than bagging or dagging.

Meta-Decision Trees (MDTs) was introduced in [20]. MDT leaves specify which classifier should be used to obtain a prediction instead of giving a prediction. A new algorithm for learning MDTs based on C4.5 Ordinary Decision Tree (ODT) learning algorithm was presented. Classifiers generated by five different learning algorithms comprising two decision trees, a rule learning algorithm, a nearest neighbour algorithm as well as Naive Bayes algorithm were combined. The results showed that Stacking with MDTs performs better than voting and stacking with ODTs.

Vasileios et al carried out evaluation of data mining algorithms on molecular dynamic trajectories. 65 classifiers available in the well-known data mining toolkit-Waikato Environment for Knowledge Analysis (WEKA) were used for this experiment. Judgement was based on the classification errors. The result showed that: Meta classifiers outperform other groups when applied to molecular dynamic datasets, Random forest and Rotation forest are the best classifiers for all the three datasets and finally, classification by clustering gave the highest classification error [15].

Performance study of Rule-Based Classification Techniques was done across multiple Database Relations in [16]. Four well-known classification techniques namely; Decision Tree, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Partial Tree (PART) and RIDOR were applied on three real multi-relational databases using WEKA. The results of this experiment show that:

1. PART and Decision Tree algorithms perform better than the others if the number of relations of the dataset is greater than 20.
2. PART and RIPPER do better if the number of tuples of the dataset is greater than 25,000.
3. PART, RIPPER shows better performance if the dataset's number of attributes is greater 25.
4. Decision Tree, PART if the dataset's number of foreign keys is greater than 2.

Classification results of two models i.e. Random Forest and the J48 were done on 20 versatile datasets in [2]. The classification results showed that Random forest gave better results for the same number of attributes and large datasets, while J48 is convenient with small data sets. Model combination was compared with data combination in [18]. The results showed that model combination performs better than data combination when the batches are taken at random from a single data source and the same learning method is applied on each.

Comparison of Decision Tree methods for finding active objects was presented in [24]. Reduced Error Pruning (REP) Tree, Random Tree, Decision Stump, Random Forest, J48, Naïve Bayes (NB) Tree and Alternating Decision (AD) Tree all available in WEKA were considered. The experimental results showed that when discriminating active objects from non-active objects, AD Tree performs better when accuracy is considered, Decision stump performs better when

speed is considered and J48 is a better choice when both accuracy and speed are considered.

The essence of these performance comparisons is to guide on the use of appropriate data mining technique or algorithm that can lead to optimal performance which will be of immense advantage to the data miners. This work also falls in this category with primary focus on obtaining optimal performance from Stacked Generalization algorithm using malaria fever datasets.

3. The Research Method

3.1. Research and Review

Analysis of data mining methods was carried out and different methods of combining multiple models were keenly studied. The concept of Stacked Generalization and Meta-learning were carefully looked into. Some related works on comparison of data mining techniques were considered. Malaria fever was reviewed and consultations with medical experts on clinical diagnosis of malaria fever were done.

3.2. Data Collection and Description of Data Sets

Data on Malaria fever cases were collected from Adetoyin Hospital, Ado-Ekiti, Nigeria and Afe Babalola University Health Centre, Ado-Ekiti, Nigeria for a period of six months by direct observation of the medical records of malaria fever patients using purposive sampling technique. One thousand two hundred and twenty five (1225) instances of the data were used to train the models while Four hundred and Eight (408) instances were used as test set. The observed nineteen symptoms (conditional attributes): Weakness, Abdominal Pain, Cough, Body Pain, Fever, Rigour, Cold, Anorexia, Headache, Catarrh, Insomnia, Yellow Urine, Vomiting, Joint Pain, Dizziness, Ill-looking, Convulsion, Temperature and Diarrhea were put into consideration. Each conditional attribute was assigned a value from (High, Low, None) based on patients feeling. Using the level of severities of the available conditional attributes (symptoms) of each patient, a record corresponding to the patient is assigned a class from (Very High, High, Moderate, Low and Very Low) by medical experts.

3.3. Stacked Generalization Method

Bootstrapped samples of the training data are used to train an ensemble of classifiers to create the base learners or level learners or base classifiers; this is the phase 1 of Stacked Generalization. The outputs from the level learners are then used to train a Meta-learner which is also a classifier and the output of the Meta-learner is the final output of the stacked Generalization method. The training of a Meta-learner using the outputs from the level learners as the training set is the phase 2 of the Stacked Generalization approach. The purpose of the Stacked Generalization is to learn if the training data have been properly learned by the base classifiers during phase 1. If there is any region of the training data that is not properly learned by a particular base classifier and it consistently makes wrong prediction on the region. The metal learner (Phase2) may learn this behaviour and combining this with the learned behaviours of other classifiers may do the correction and thus make a correct prediction. Phase1 of the stacked generalization approach employs cross validation method in which the entire data is divided into P blocks and all the base classifiers are trained independently on P-1 blocks of the training data. Each classifier is tested on the P^{th} block of the data not seeing during the training. The outputs of the base classifiers on the P^{th} block combined with the actual correct labels for those blocks served as the training set for the Meta-learner in the phase 2 [11].

In this work, enhanced version of the Stacked Generalization Algorithm was used to carry out the diagnosis of malaria fever. Stacked Generalization was first proposed by Wolpert in 1992 and later enhanced by Ting and Witten [18].

3.3.1. Stacked Generalization Approach to Malaria Diagnosis

Decision Tree Algorithms-J48, Random Tree, REP Tree and Rule Based Classification Techniques- JRIP, PART, RIDOR were used as the base learners. Decision Tree uses the information gain,

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

to determine the splitting Node (N) which represents the tuples of partition D, where p_i is the probability that an arbitrary tuple in D belongs to Class C_i and is estimated by $|C_i, D|/|D|$ while the Rule Based Classification Techniques use the modified information gain (G) for classification:

$$G = p\left[\log\left(p/t\right) - \log\left(P/T\right)\right] \quad (2)$$

Where p is true positives, t is total example covered by the rule, P is number of positive examples of this class and T is total number of examples of this class.

A malaria fever diagnosis model was built using an adapted enhanced Stacked Generalization Algorithm that combines the six base learners. Given a malaria fever data set

$$S = \{(p_n, q_n), n = 1,\ldots,N\}, \quad (3)$$

Where p_n is the class value and q_n represents the attribute values of the nth instance of malaria fever dataset, randomly split the data into J almost equal parts $D_1,\ldots, D_J$. Define D_j and $D^{(-j)} = D\text{-}D_j$ to be the test

and training sets respectively for the *jth* fold of a J-fold cross-validation. Given K learning algorithms (Random Tree, REP Tree, J48, JRIP, PART and RIDOR), called level-0 generalizers, invoke the kth algorithm on the data in the training set $D^{(-j)}$ to induce a model $L_k^{(-j)}$, for k = 1, 2, 3,…6.

Let the output from the Level-0 models be a set of class probabilities rather than a single class prediction. If model $L_k^{(-j)}$ is used to classify an instance q in D_j, Let $B_{ki}^{(-j)}(q)$ denote the probability of the *ith* output class :

$$R_{kin} = B_{ki}^{(-j)}(q_n) \quad (4)$$

The class probabilities are assembled with the original class to constitute Level-1 data (LDD):

$$L_{DD} = \{(p_n, R_{1In,\ldots}, R_{1In},\ldots,R_{klm,\ldots}, R_{KIn},\ldots,R_{KIn}),\ n=1,\ldots,N. \quad \textit{(assuming there are I classes).} \quad (5)$$

NNGE/Forest Tree learning strategy called level-1 generalizer was finally used on this training data (L_{DD}) to produce a level-1 model M_f whose output is the final classification result for malaria fever instances.

3.3.2. Algorithms Exploited in this Work

This work was carried out using the Waikato Environment for Knowledge Analysis (WEKA). WEKA is a world known and accepted data mining tool which implemented a number of machine learning and data mining algorithms. It is openly available for both academic research and industrial use. "Applications written using the Weka class libraries can be run on any computer with a Web browsing capability" [6]. All the algorithms exploited in this work are available in WEKA. These algorithms are described briefly below.

- *REP Tree*: Reduced Error Pruning Tree (REP Tree) is a fast decision learner. It uses information gain as the splitting criterion to build decision or regression trees. REP Tree uses reduced error pruning method for pruning. Further details on REP Tree can be gotten from [2, 12, 24].
- *Random Tree*: Random Tree is one of the promising machine learning algorithms. Here a tree is selected at random from a set of possible trees. All the trees have equal probability of being selected. Models with improved accuracy can be obtained by combing a large set of random trees [2, 12, 24]. Please consult [2, 12, 24] for more details.
- *J48*: J48 is an enhanced version of C4.5 algorithm. It uses recursive partitioning of data to generate decision trees for a given dataset, Depth-first strategy for decision growing and information gain for slitting. Post-pruning approach is used to avoid over fitting. A full tree is allowed to be grown by the algorithm, after which the removal of the branches that lead to over fitting. For further reading, see [24].
- *Random Forest*: Random forest uses meta-learning method to combine models from trees obtained from different random samples obtained from a large dataset. The trees generated from random samples of the data set are not pruned. Classification or regression trees are first generated, after which they are ensemble for making prediction. During aggregation, it uses majority vote for classification or averaging for regression [24].
- *JRIP*: JRIP classifiers a new instance (element) by using propositional rules generated after the training. It can be categorised as an inference and rule-based learner [14]. All positive examples are covered by its rule set and the JRIP algorithm does very well on large and noisy datasets. It uses both IREP (Incremental Reduced Error Pruning) and RIPPER strategies. The first rule is generated using IREP and later RIPPER. JRIP partitions the training data set into two subsets, usually 2/3 of the training set is used as the growing set while the rest 1/3 are used as the pruning set. The rules are generated from the growing set [7]. JRIP Algorithm is available in [21].
- *PART*: Partial Tree (PART) uses two common data mining strategies. It employs the divide-and-conquer strategy for decision tree learning and the separate-and-conquer strategy for the rule learning and uses separate-and-conquer strategy to build the rules. This promising classifier avoids global optimization step being used by C4.5 and RIPPER [16]. PART detailed algorithm can be found in [16].
- *RIDOR*: Default rule is generated first by RIDOR (Ripple Down Rule Learner), after which the exceptions for the default rule with the least (weighted) error rate are generated. Best exception rules are generated for each exception and iteration continues until no exception remains. Thus, it carries out a tree-like expansion of exceptions and the leaf has only default rules but no exceptions. The exceptions constitute a set of rules that predict the improper instances in default rules [7, 16]. Details of RIDOR algorithm are available in [7].
- *NNGE*: Non-Nested Generalized Exemplars (NNGE) is a classification technique which works in a manner similar to the Nearest-Neighbour algorithm. It employs non-nested generalized exemplars, which are hyper rectangles that can be viewed as if then rules [3, 7]. NNGE uses Euclidean distance function to classify a new example by calculating the nearest neighbour in the exemplar/hyper rectangle database. Further details on NNGE and its algorithm can be accessed from [4].

4. Experimental Setup and Results

For the ease of data mining, programming and the fact that it is easier to work around with numbers, decision attribute values (Classes)-Very Low, Low, Moderate, High and Very High were thus converted to integer numbers 1,2,3,4 and 5 respectively. The conditional attribute values High, Low and None were also converted to the integer numbers 2, 1, 0 respectively. For example, HEC = 2 means the symptom headache is perceived to be high and MAL = 5 means Malaria is diagnosed to be Very High. After both the training and testing sets were formatted into acceptable format, classification experiments were then carried out. Experimental set up was broken into five major steps described below:

- *Step* 1: A predictive model was generated by each of the six base learners using the one thousand, two hundred and twenty five (1225) training set. The six different models generated were individually tested on both the training set and the four hundred and eight (408) testing set. The results obtained are given in the Tables 1 and 2 respectively, while classification accuracy by class on both the training and the testing sets are displayed in Table 3 and Table 4 respectively.

Table 1. Classification accuracy of the six base learners on the training set.

Method	C	IN	% C	% IN
PART	1225	0	100	0
REP Tree	1225	0	100	0
J48	1225	0	100	0
Random Tree	1225	0	100	0
RIDOR	1189	36	97.0612	2.9388
JRIP	1213	12	99.0204	0.9796

Table 2. Classification accuracy of the six base learners on the testing set.

Method	C	IN	% C	% IN
PART	400	8	98.0392	1.9608
REP Tree	400	8	98.0392	1.9608
J48	400	8	98.0392	1.9608
Random Tree	400	8	98.0392	1.9608
RIDOR	392	16	96.0784	3.9216
JRIP	384	24	94.1176	5.8824

Table 3. The six base learners' classification accuracy by class on the training set.

	V.HIGH		HIGH		MODERATE		LOW		V.LOW	
METHOD	**C**	**IN**	**C**	**IN**	**C**	**IN**	**C**	**IN**	**C**	**IN**
PART	134	0	635	0	247	0	135	0	74	0
REP Tree	134	0	635	0	247	0	135	0	74	0
J48	134	0	635	0	247	0	135	0	74	0
Random Tree	134	0	635	0	247	0	135	0	74	0
RIDOR	134	0	611	24	235	12	135	0	74	0
JRIP	134	0	635	0	247	0	123	12	74	0

Table 4. The six base learners' classification accuracy by class on the testing set.

	V.HIGH		HIGH		MODERATE		LOW		V.LOW	
METHOD	**C**	**IN**	**C**	**IN**	**C**	**IN**	**C**	**IN**	**C**	**IN**
PART	33	8	258	0	42	0	49	0	18	0
REP Tree	41	0	258	0	34	8	49	0	18	0
J48	41	0	258	0	34	8	49	0	18	0
Random Tree	41	0	258	0	34	8	49	0	18	0
RIDOR	33	8	258	0	34	8	49	0	18	0
JRIP	41	0	242	16	34	8	49	0	18	0

NOTE: "C" means Correctly Classified, "IN" means Incorrectly Classified

- *Observations from Step* 1: It was observed from the Table 1 and Table 2 that PART, REP TREE, J48 and Random Tree classified all the training sets correctly attaining 100% in each case. JRIP misclassified 12 instances while RIDOR misclassified 36 instances which make it the least of the six in terms of accuracy. None of the six base classifiers has 100% performance on the testing test. PART, REP Tree, J48 and Random Tree all misclassified eight (8) instances each; RIDOR misclassified 16 instances while JRIP misclassified 32 instances. Classification accuracy of the least classifier which gives 94.1176% may be adjudged excellent in some situations but there should be better care for performance when it comes to health issues, then the need to work further to improve on the results. It was also evident in Table 3 and Table 4 that each algorithm has its own weakness(s). For instance PART, REP Tree, J48 and Random Tree each has 98.0392% detection rate on the testing set and each misclassified eight (8) instances but the classes misclassified differ. PART misclassified Very High while REP Tree, J48 and Random Tree misclassified Moderate. JRIP with lower overall performance to the leading first four even classified all the forty one (41) Very High instances correctly. Thus, ensemble of these base classifiers may produce a better performance since they show instability (i.e., makes errors in different part of the training set).

- *Step* 2: The six base learners (base level algorithms) were arranged in two different ways to form two cases:
  - *Case* 1: PART, REP Tree, J48, Random Tree, RIDOR, JRIP
  - *Case* 2: JRIP, PART, REP Tree, Random Tree, RIDOR, J48

- *Step* 3: Stacked Generalization algorithm was used to combine the six base learners in case1 using Random Forest and Non-Nested Generalized Exemplars (NNGE) as Meta-Learners differently. The two resulting ensemble models were each tested on both the training set (1225 instances) and testing

set (408 instances). The results obtained from step3 are displayed in Table 5.

- *Observation from Step* 3: It was observed from the Case1 results that Stacked Generalization gives a better performance over any of the single base level algorithms, attaining 100% for both the training and testing sets i.e., all the 1225 training instances and 408 testing instances were correctly classified.
- *Step* 4: Stacked Generalization algorithm was used to combine the six base learners in Case II using Random Forest and NNGE as Meta Classifiers differently. The two resulting Stacked Generalization models were each tested on both the training set (1225 instances) and the testing set (408 instances). The results obtained are displayed in Table 6 and 7.

Table 5. Classification accuracy of the training set for case 1 using Random Forest and NNGE as meta-classifiers separately.

	Training Set				Testing Set			
Meta-Learner	**C**	**IN**	**%C**	**%IN**	**C**	**IN**	**%C**	**%IN**
Random Forest	1225	0	100	0	408	0	100	0
NNGE	1255	0	100	0	408	0	100	0

Table 6. Classification accuracy of the training set for case 2 using Random Forest and NNGE as meta-classifiers separately.

	Training Set				Testing Set			
Meta-Learner	**C**	**IN**	**%C**	**%IN**	**C**	**IN**	**%C**	**%IN**
Random Forest	1225	0	100	0	400	8	98.0392	1.9608
NNGE	1255	0	100	0	408	0	100	0

Table 7. Classification accuracy by class of testing set for case 3 using random forest and NNGE as meta-learners separately.

	V.HIGH		HIGH		MODERATE		LOW		V.LOW	
Meta-Learner	**C**	**IN**	**C**	**IN**	**C**	**IN**	**C**	**IN**	**C**	**IN**
Random Forest	33	8	258	0	42	0	49	0	18	0
NNGE	41	0	258	0	42	0	49	0	18	0

NOTE: "C" means Correctly Classified, "IN" means Incorrectly Classified.

- *Observations from Step* 4 : The results obtained from the Case 2 Experiment show that NNGE outperforms Random Forest as a Meta Learner. NNGE as a Meta Learner gives 100% on the training set and also 100% on the Testing set while Random Forest as a Meta Learner gives 100% on the training set but 98.0392% on the testing set, misclassified a total of eight (8) instances of the Class "Very High" as showing Table 7.
- *Step* 5: Comparison of the results from Step1, Step 3 and Step 4 and selection of the best Ensemble Learning classification Model that gives the highest performance accuracy.

The results obtained from the Step1, Step3 and Step4 were keenly studied, leading to the following conclusions.

5. Summary of the Experimental Setup and Results

1. The six base learners are unstable since their performances are not the same and there is evidence of misclassified different classes.
2. None of the six base learners gives 100% classification accuracy on the Testing Set (J48= 98.0392%, Random Tree=98.0392%,REP Tree=98.0392% JRIP =94.1176%, RIDOR= 96.0784% , PART=98.0392%).
3. Stacked Generalization approach outperforms any of the six base learners, gives 100% on Training Set and 100% on the Testing Set (Step 3 and Step 4 results).
4. NNGE as a Meta Learner outperforms Random Forest as a Meta Learner (Step 3 and Step 4).
5. The order of arrangement of the Base Learners may affect the output of Stacked Generalization though subject to the choice of the Meta Learner. Random Forest as a Meta Learner gives 100% on Testing Set for Case I but 98.0392% for Case II whereas NNGE gives 100% on Training Set and also 100% on the Testing Set for Case 2.
6. From (5) above, performance of NNGE as Meta Learner is independent of the order of arrangement of the Base Learners whereas that of Random Forest as a Meta Learner is dependent of the order of arrangement of the Base Learners.
7. Finally, Stacking or Stacked Generalization of the Six Base Learners (J48, Random Forest, REP Tree, PART, RIDOR , JRIP) with NNGE as a Meta Learner gives the best performance accuracy.

6. Conclusions

Data mining has been applied successfully to different problems of the society either as predictive or descriptive models. It is evident from literature that the various available data mining algorithms have their strengths and weaknesses and thus combination of two or more algorithms or models could give a better performance. Stacked Generalization is a method of combining multiple classifiers referred to as base learners for a better accuracy. This study was carried out using the world known Data mining tool kit (Waikato Environment for Knowledge Analysis (WEKA)). Though, Stacked Generalization has been proved to be effective when the base learners are unstable but may not give an optimal performance if the base learners and the meta- learner are not carefully selected. The study further shows that the order of arrangement of the base learners may have effect on the performance of the Stacked Generalization method. It further reveals that the choice of meta-learner can also determine the performance of Stacked Generalization. NNGE outperforms Forest Tree as a meta-learner and its performance is independent of the

order of arrangement of the base learners as against Forest Tree. It is therefore recommended that Data Miners should take note of these observations in their future work.

References

[1] Abbasimehr H., Sefak M., and Tarokh M., "A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction," *The International Arab Journal of Information Technology*, vol. 11, no. 6, pp. 599-606, 2014.

[2] Ali J., Khan R., and Ahmad N., "Random Forests and Decision Tree," *International Journal of Computer Science Issues*, vol. 9, no. 5, 2012.

[3] Datta R. and Saha S., "An Empirical Comparison of Rule Based Classification Techniques in Medical Databases," *Indian Institute of Foreign Trade*, 2011.

[4] Devasena L., Sumathi T., Gomathi V., and Hemalatha M., "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set," *Bonfring International Journal of Man Machine Interface*, vol. 1, pp. 5-9, 2011.

[5] Fan D., Chan K., and Salvatore J., "A Comparative Evaluation of Combiner and Stacked Generalization," *in Proceedings of AAAI-96 Workshop on Integrating Multiple Learned Models*, pp. 40-46, 1996.

[6] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I., "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10-18, 2009.

[7] Karahoca A., Karahoca D., and Aydin N., "Benchmarking the Data Mining Algorithms with Adaptive Neuro-Fuzzy Inference in System GSM Churn Management," *Data Mining and Knowledge Discovery in Real Life Applications*, Vienna, pp. 230-241, 2009.

[8] Keleş A., "Expert Doctor Verdis: Integrated Medical Expert System," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 22, pp. 1032-1043, 2014.

[9] Mahajan A. and Ganpati A., "Performance Evaluation of Rule Based Classification Algorithms," *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 3, no. 10, pp. 3546-3550, 2014.

[10] Parkhi V., Pawar P., and Surve A., "Computer Automation for Malaria Parasite Detection using Linear Programming," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 5, pp. 1984-1988, 2013.

[11] Polikar R., *Ensemble Learning*, Springer, 2008.

[12] Ravichandran S., Srinlvasan V., and Ramasamy C., "Comparative Study on Decision Tree Techniques for Mobile Call Detail Record," *Journal of Communication and Computer*, vol. 9, pp. 1331-1335, 2012.

[13] Soomro A., Memon N., and Menopn M., "Knowledge Based Expert System for Symptomatic Automated HealthCare," *Sindh University Research Journal (Science Series)*, vol. 43, no. 1-a, pp. 79-84, 2011.

[14] Tarun I., Gerardo B., and Tanguilig B, "Generating Licensure Examination Performance Models Using PART and JRIP Classifiers: A Data Mining Application in Education," *International Journal of Computer and Communication Engineering*, vol. 3, no. 3, pp. 203-207, 2014.

[15] Tatsis V., Tjortjis C., and Tzirakis P., "Evaluating Data Mining Algorithms using Molecular Dynamics Trajectories," *International Journal Data Mining and Bioinformatics*, vol. 8, no. 2, pp. 169-187, 2013.

[16] Thangaraj M. and Vijayalakshmi C., "Performance Study on Rule-Based Classification Techniques across Multiple Database Relations," *International Journal of Applied Information Systems*, vol. 5, no. 4, pp. 1-7, 2013.

[17] Ting K. and Witten I., "Stacking Bagged and Dagged Models," *in Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, pp. 367-375, 1997.

[18] Ting K. and Witten I., "Stacked Generalization: When does it Work?," *in Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pp. 866-871, 1997.

[19] Ting K., Low B., and Witten I., "Learning from Batched Data: Model Combination vs Data Combination," *Knowledge and Information Systems*, vol. 1, no.1, pp. 83-106, 1999.

[20] Todorovski L. and Džeroski S., "Combining Classifiers with Meta Decision Trees," *Machine Learning*, vol. 50, no. 3, pp. 223-249, 2003.

[21] Veeralakshmi V. and Ramyachitra D., "Ripple Down Rule Learner (RIDOR) Classifier for IRIS Dataset," *International Journal of Computer Science Engineering*, vol. 4, no. 3, pp. 79-85, 2015.

[22] Vilalta R., Giraud-Carrier C., Brazdil P., and Soares C., "Using Meta-Learning to Support Data Mining," *International Journal of Computer Science and Applications*, vol. 1, no. 1, pp. 31-45, 2004.

[23] Williams S. and Olatunji O., "Hybrid Intelligent System for the Diagnosis of Typhoid Fever," *Journal of Computer Engineering and Information Technology*, vol. 2, no. 2, pp. 1-9, 2013.

[24] Zhao Y. and Zhang Y., "Comparison of Decision Tree Methods for Finding Active Objects,"

Advances of Space Research, vol. 41, no. 12, pp. 1955-1959, 2007.

Abiodun Oguntimilehin is a Senior Lecturer in the Department of Computer Science, Afe Babalola University, Nigeria. He obtained Ph.D in Computer Science from the Federal University of Technology, Akure, Nigeria. He is a chartered member, Computer Professionals (Registration Council of Nigeria) and Nigeria Computer Society. His research interests are Medical Informatics, Artificial Intelligence and Machine Learning. He has a number of publications in both reputable local and international journals.

Olusola Adetunmbi is a Professor in the department of Computer Science, Federal University of Technology, Akure, Nigeria. He obtained a Ph.D degree in Computer Science from the Federal University of Technology, Akure, Nigeria. His Research interests are Information Security, Machine Learning and Natural Language Processing. He is a member of IEEE Computer Society, International Studies on Advanced Intelligence, Computer Professionals (Registration Council of Nigeria) and Nigeria Computer Society. He has a number of publications in both reputable local and international journals.

Innocent Osho is a Professor of Veterinary Parasitology and ethno-veterinary medicine in the Department of Animal Production and Health, Federal University of Technology, Akure, Nigeria. He obtained PhD in Animal Parasitology from the Federal University of Technology, Akure, Nigeria. He is a member of Nigeria Veterinary Medicine Association and European Phyto-chemical Association among others. He has over 63 publications in some reputable local and international journals.